DOCUMENT RESUME

ED 044 433 TM 000 148

AUTHOR Hambleton, Ronald K.; Traub, Ross E.
TITLE Information Curves and Efficiency of Three Logistic Test Models.
INSTITUTION Massachusetts Univ., Amherst. School of Education.
PUB DATE Mar 70
NOTE 19p.; Paper presented at the annual meeting of the American Psychological Association, Miami, Florida, March 1970

EDRS PRICE EDRS Price MF-$0.25 HC-$1.05
DESCRIPTORS *Ability Identification, Data Collection, *Models, *Scoring, Simulation, Test Validity

ABSTRACT

The purpose of this study was to determine the efficiency of the estimates of ability provided by the one-parameter logistic model as compared to the estimates provided by the more general two- and three-parameter models. Several tests were simulated with item parameters meeting the assumptions of either the two- or three-parameter model. For each test, the information provided by ability estimates appropriate to the one-, two- and three-parameter models was compared at several ability levels. The results indicate that it is particularly important, when guessing affects test scores, to use the scoring system of the three-parameter model for estimating the ability of low-ability examinees. (Author)

Technical Reports

No. 4

INFORMATION CURVES AND EFFICIENCY OF THREE LOGISTIC TEST MODELS

Ronald K. Hambleton
University of Massachusetts

and

Ross E. Traub
The Ontario Institute for Studies in Education

March, 1970

CENTER FOR EDUCATIONAL RESEARCH

School of Education
University of Massachusetts
Amherst

INFORMATION CURVES AND EFFICIENCY OF THREE LOGISTIC TEST MODELS[1]

Ronald K. Hambleton
University of Massachusetts

and

Ross E. Traub
The Ontario Institute for Studies in Education

[1]Paper presented at the annual meeting of the American Psychological Association, Miami, 1970.

INFORMATION CURVES AND EFFICIENCY OF THREE LOGISTIC TEST MODELS

Ronald K. Hambleton
University of Massachusetts

and

Rose E. Traub
The Ontario Institute for Studies in Education

One way of evaluating a latent trait model for tests is in terms of the precision with which it estimates an examinee's ability: The more precise the estimate, the more information the model can be said to provide. Birnbaum (1968) operationalized this conception of information as the quantity inversely proportional to the squared length of the confidence interval for the estimate of an examinee's ability. Defined in this way, the amount of information in a test is a function of ability. Mathematically, Birnbaum's information function may be defined as

$$I[\theta,x] = \left[\sum_{g=1}^{n} w_g^2 P_g(\theta) Q_g(\theta)\right]^{-1}\left[\sum_{g=1}^{n} w_g P_g'(\theta)\right]^2 . \qquad (1)$$

In equation (1), I is the amount of the information at ability level θ provided by scoring formula, x, where

$$x = \sum_{g=1}^{n} w_g u_g ; \qquad (2)$$

n is the number of items in the test, w_g is the scoring weight for item g, and u_g is a function which takes the value one if item g is answered correctly, and zero otherwise. The remaining terms of

equation (1) are defined as follows:

$$P_g(\theta) = c_g + (1 - c_g)\,[1 + e^{-Da_g(\theta - b_g)}]^{-1}\,, \tag{3}$$

$$Q_g(\theta) = 1 - P_g(\theta)\,, \tag{4}$$

and

$$P'_g(\theta) = \frac{\partial P_g(\theta)}{\partial \theta}\,. \tag{5}$$

$P_g(\theta)$ is the characteristic curve for item g with its mathematical form specified by the test model; it gives the probability that an examinee of ability θ answers item g correctly. In the three-parameter logistic model, (Birnbaum, 1968), the item characteristic curve takes the form presented in equation (3). The parameters b_g and a_g are usually referred to, respectively, as the index of difficulty and discrimination of item g, while parameter c_g, the lower asymptote of the item characteristic curve, may be thought of as the guessing parameter. The constant D is a scaling factor that is usually chosen to be 1.7 to make the logistic distribution function conform as closely as possible to the normal (Lord, 1952) A two-parameter logistic model (Birnbaum, 1957; 1958a; 1958b; 1968) may be obtained from the three-parameter model by assuming that the effect of guessing on test scores is negligible and setting c_g in equation (3) to zero. If, in addition, it is assumed that the items in a test have equal discriminating power (i.e., $a_g = \bar{a}$ for all g, $g = 1, 2, \ldots, n$) the resulting item characteristic curve has but one free parameter per item (i.e. b_g) and specifies a model that can be shown to be formally equivalent to a test model developed by

Rasch (1960; 1966).

Birnbaum (1968, p. 454) demonstrated that the maximum value of $I[\theta, x]$, represented as $I[\theta]$, is given by

$$I(\theta) = \sum_{g=1}^{n} \left\{ [P'_g(\theta)]^2 \Big/ P_g(\theta)\, Q_g(\theta) \right\} \qquad (6)$$

In general, $I[\theta, x] \leq I[\theta]$. Equality holds when the scoring weights, w_g, are chosen such that

$$w_g = P'_g(\theta) \,/\, P_g(\theta)\, Q_g(\theta) \;,\; g = 1, 2, \ldots, n \;, \qquad (7)$$

except for a possible scaling factor. Thus to maximize the information function and consequently minimize the width of the confidence band about an ability estimate under the one-, two- and three-parameter logistic models, the scoring weights should be chosen to be 1, Da_g, and $Da_g\, \Psi[Da_g(\theta - b_g) - \log c_g]$, $g = 1, 2, \ldots, n$, respectively. (In the third weight, Ψ is the logistic distribution function.) Notice that only in the case of the three-parameter model are the weights dependent on ability. The scoring system of the three-parameter model has the effect of reducing the weight assigned to correct answers on items with a sizeable guessing parameter. Moreover, the weight for such items is smallest for low ability examinees who are most likely to have answered by guessing, and becomes increasingly large as the ability of the examinee increases.

If scoring weights different from the optimal weights specified by a test model are used, the information derived by using these inappropriate weights to score a test will be less than what is potentially available. Birnbaum used the term efficiency to refer to the information

lost due to the use of less than optimal scoring weights. The concept of efficiency may be formally explicated as follows. Assuming a particular test model is the true model, let $\underline{I}_1[\underline{\theta}, \underline{x}_1]$ and $\underline{I}_2[\underline{\theta}, \underline{x}_2]$ represent the information functions of any two scoring formulas $\underline{x}_1$ and $\underline{x}_2$ respectively. Then, the ratio $\underline{I}_1[\underline{\theta}, \underline{x}_1] / \underline{I}_2[\underline{\theta}, \underline{x}_2]$ is called the relative efficiency (at $\underline{\theta}$) of $\underline{x}_1$ to $\underline{x}_2$. If the scoring weights used in $\underline{x}_2$ are such that $I_2(\hat{\theta}) = I_2[\theta, x_2]$, then the ratio of $\underline{I}_1[\underline{\theta}, \underline{x}_1] / \underline{I}_2(\underline{\theta})$ is called the efficiency (at $\underline{\theta}$) of $\underline{x}_1$. Thus, it is possible, using the optimal scoring weights specified by a model, to investigate the relative efficiency of the model at estimating ability when a test is known to be composed of items that conform to the assumptions of a more general model. For example, the one-parameter logistic (Rasch) model specifies unit scoring weights for estimating ability. The efficiency of scores based on these weights when the items in a test conform to the assumptions of a two-parameter logistic model is given by

$$\mathrm{Eff}[\theta, x_1] = \frac{I[\theta, x_1]}{I(\theta)} = \frac{\left[\sum_{g=1}^{n} P'_g(\theta)\right]^2}{\left[\sum_{g=1}^{n} P_g(\theta)\, Q_g(\theta)\right]\left[\sum_{g=1}^{n} Da_g\, P'_g(\theta)\right]}, \qquad (8)$$

where $$x_1 = \sum_{g=1}^{n} u_g, \qquad (9)$$

and

$$P_g(\theta) = \left[1 + e^{-Da_g(\theta - b_g)}\right]^{-1}, \qquad (10)$$

The efficiency of scores computed from the weights specified by the two-parameter logistic model when the items of a test conform to the assumptions of the three-parameter model is given by

$$\mathrm{Eff}[\theta, x_2] = \frac{I[\theta, x_2]}{I(\theta)} = \frac{\left[\sum_{g=1}^{n} Da_g P'_g(\theta)\right]^2 \left[\sum_{g=1}^{n} D^2 a_g^2 \psi^2 [Da_g(\theta - b_g) - \log c_g]\, P_g(\theta) Q_g(\theta)\right]}{\left[\sum_{g=1}^{n} D^2 a_g^2 P_g(\theta) Q_g(\theta)\right] \left[\sum_{g=1}^{n} Da_g \psi [Da_g(\theta - b_g) - \log c_g]\, P'_g(\theta)\right]^2},$$

where $x_2 = \sum_{g=1}^{n} Da_g u_g$, (12)

and $P_g(\theta)$ is defined as in equation (3).

The question of efficiency has been considered in at least two previous studies. Birnbaum (1968) investigated the efficiency of unit scoring weights when the weights specified by the two-parameter model were appropriate. He did this for abilities in the range $-3 \leq \theta \leq 3$ while systematically varying the range of the distribution of discrimination parameters. Birnbaum considered some tests in which the discrimination parameters of the items were located half at one end of the range of the distribution of discrimination parameters, half at the other end. The items in Birnbaum's tests were all of middle difficulty, that is $b_g = 0$, $g = 1, 2, \ldots, n$. When there was a small difference between the two possible values of the discrimination index (0.44 vs. 0.58), efficiency was about 97%. When the values of the discrimination index were 0.31 or 0.75, efficiency was reduced, and varied from about 80% to about 90% depending on the level of ability. When the two values of the discrimination parameter were made to approximate the maximum difference

that is observed in practice (0.20 vs. 0.98), efficiency varied from about 60% to about 70%, again depending on ability. Birnbaum also considered the more typical case in which the items of a test have discrimination parameters distributed more or less uniformly across the range 0.20 to 0.98. In this case, efficiency was about 80%. Using a scoring system with an efficiency of 80% is equivalent to discarding 1/5 of the information available in the test. Clearly, in such instances it would be inefficient to use unweighted test scores.

Lord (1968) investigated the efficiency of ability estimates based on unit scoring weights when optimal estimates would be based on the weights specified by the three-parameter logistic model. He found that the efficiency of unit-weight scores on the verbal part of the scholastic aptitude test where it was assumed that the three-parameter model was the true model varied from 55% at the lowest ability level to a maximum of 90% at a high ability level. Here again, the importance was demonstrated of using scoring weights appropriate to a more general test model.

Purpose

Recently, there has been increased interest in logistic test models, particularly the one-parameter logistic (Rasch) model. Because the restrictive assumptions of the one-parameter model are often violated by test data [see Hambleton (1969) for a summary of the evidence] the model will usually not fit data as well as the more general logistic models. Hence, using the one-parameter model to estimate ability when a more general model would provide a more appropriate estimate will result in a loss of information in the sense defined earlier.

The questions asked in this study were as follows: How much information is obtained about an examinee's ability using the scoring systems of the one-, two- and three-parameter logistic test models as the range of the distribution of item discrimination parameters and the mean level of guessing on the items are varied systematically in simulated tests? Under these circumstances, what is the efficiency of the scoring systems of the less appropriate one- and two-parameter models when the comparative standard is the amount of information provided by the more appropriate two- and three-parameter models? Since information curves and efficiency are both a function of ability, answers to the two questions were obtained for different values of θ.

Methodology

Generation of Item Parameters

To begin with, it was assumed that only a single latent ability was being measured. This is an assumption typically made in latent trait theory (McDonald, 1967). The situation which was envisioned as being in some sense typical of nature was one in which scores on this single latent ability are normally distributed in the population. A suitable scaling of the ability continuum would establish a mean of the ability distribution of zero and a standard deviation of one. Under these conditions, over 99% of the population would have ability scores on the interval [-3, 3]. These limits for the range of ability were chosen for the study.

Tests were simulated so that the items ranged in difficulty within reasonable limits for the group being tested. In effect, it was assumed the test would contain no item so easy that more than 95% of a

population approximately normally distributed on the interval [-3, 3] would get it correct; also, no item would be so difficult that less than 5% of the population would get it correct. Difficulty parameters $\underline{b}_g$, $g = 1, 2, \ldots, n$, were randomly assigned to each of the items in a simulated test subject to the restriction they were drawn from a population distribution of difficulty parameters that was rectangular on the interval [-2, 2] with a mean of zero. Lord's (1968) work reveals this choice of assumed distribution and range of item difficulty parameters to be realistic, at least for the kind of test he studied.

The item discrimination parameters, $\underline{a}_g$, $g = 1, 2, \ldots, n$, were assumed to be drawn from a uniform population distribution with a mean of 0.59 and a range which was systematically varied across simulations between zero and 0.80, inclusive. The results obtained by Lord (1968) and Ross (1966) support the choice of this form of distribution for the discrimination parameters.

The magnitude of the item guessing parameters, $\underline{c}_g$, $g = 1, 2, \ldots, n$, for each set of test data was controlled by the value of $\bar{\underline{c}}$, where $\bar{\underline{c}}$ was the mean of the guessing parameters of the items in a simulated test. Assuming five-option multiple-choice tests and a heterogeneous ability group, it seemed reasonable also to assume that individual values of $\underline{c}_g$ and $\bar{\underline{c}}$ would be bounded on the interval [.00, .20]. Give a specified value of $\bar{\underline{c}}$, the $\underline{c}_g$'s were generated subject to two constraints:

$$(1) \quad \bar{c} = \sum_{g=1}^{n} c_g/n$$

$$(2) \quad \bar{c} - \min\{.20 - \bar{c}, \bar{c}\} \leq c_g \leq \bar{c} + \min\{.20 - \bar{c}, \bar{c}\} , \; g = 1, 2, \ldots, n.$$

Procedure

Four ranges of the distribution of discrimination parameters were considered: 0.00, 0.20, 0.40 and 0.80. Three mean levels of the guessing parameter were considered: 0.00, 0.10 and 0.20. (In the case of $\bar{c} = 0.00$ or $\bar{c} = 0.20$, all the values of c_g were zero or 0.20, respectively.) Under the conditions specified above, item parameters were generated at random by computer for eleven of the twelve possible combinations of the range of distribution of discrimination parameters and mean level of guessing. (Excluded was the case where the range and $\bar{c}$ would be zero.) Each simulated test was assumed to have 15 items.

Taking the three-parameter logistic model to be the true model (except when $\bar{c} = 0$, in which case the two-parameter logistic model was taken to be the true model), the information provided by scores based on the weights of the one-, two- and three-parameter logistic models was computed for each of seven values of θ, $\theta = -3 + k$, $k = 0, 1, \ldots, 6$. The efficiency of the scoring systems specified by the less general test models was then determined for each level of ability. All the computations were done using a program developed by Hambleton (1970).

Results and Discussion

The notation, $x_1 = \sum_{g=1}^{n} u_g$, $x_2 = \sum_{g=1}^{n} Da_g u_g$ and $x_3 = \sum_{g=1}^{n} Da_g \Psi[Da_g (\theta - b_g) - \log c_g]$ was used for the scoring formulas specified by the one-, two- and three-parameter logistic models respectively. The quantities $I[\theta, x_1]$, $I[\theta, x_2]$, $I[\theta, x_3]$, $Eff[\theta, x_1] = I[\theta, x_1] / I[\theta, x_3]$, $Eff[\theta, x_2] = I[\theta, x_2] / I[\theta, x_3]$ and $Eff[\theta, x_1, x_2] = I[\theta, x_1] / I[\theta, x_2]$ for $\theta = -3 + k$, $k = 0, 1, \ldots, 6$, are reported in Table 1 for

eleven sets of data. When $\bar{c} = 0$, $x_2 = x_3$, hence $I[\theta, x_3]$, $Eff[\theta, x_2]$ and $Eff[\theta, x_1, x_2]$ are not reported. When the range of the discrimination parameters is zero, $x_1 = x_2$, and so $\underline{I}[\underline{\theta}, \underline{x}_2]$, $\underline{Eff}[\underline{\theta}, \underline{x}_2]$, and $\underline{Eff}[\underline{\theta}, \underline{x}_1, \underline{x}_2]$ are not reported.

The information values displayed in Table 1 reveal an approximately bell-shaped relationship between information and ability. Information is greatest near the middle of the ability distribution and much less at the extremes. When guessing occurs ($\bar{c} > 0$), less information is provided by all three scoring systems, but the decrease is particularly noticeable at low ability levels for scoring formulas $\underline{x}_1$ and $\underline{x}_2$. The relationships among the information functions of the scoring systems, under the assumption that the three-parameter model is the appropriate one, may be roughly summarized by the inequalities.

$$I[\theta, x_3] \geq I[\theta, x_2] \geq I[\theta, x_1] ,$$

This relationship appears to hold except for the situations involving very low levels of ability and $\bar{c} > 0$ when, $I[\theta, x_1] \geq [\theta, x_2]$.
It appears that when guessing is a component in test performance, unit scoring weights are better than the weights specified by the two-parameter model at estimating the ability of low ability examinees.

On additional comment should be made about information functions. It is possible to obtain any shape for the information function that is desired by judicious choice of test items (Birnbaum, 1968). The information functions described here may be considered relevant for at least some testing situations because the distributions of item parameters chosen to guide the simulation of test data were similar to what has been observed

TABLE 1

Information Curves and Efficiency

Set 1

Discrimination Parameters: $\bar{a}$ = .59 , Range = .20; .49 to .69 .
Guessing Parameters : $\bar{c}$ = .00 , Range = .00 .

Ability	$I[\theta,x_1]$	$I[\theta,x_2]$	$I[\theta,x_3]$	$Eff[\theta,x_1]$	$Eff[\theta,x_2]$	$Eff[\theta,x_1,x_2]$
-3.0	.99	.99	--	.99	--	--
-2.0	1.85	1.86	--	.99	--	--
-1.0	2.63	2.66	--	.99	--	--
0.0	2.82	2.84	--	.99	--	--
1.0	2.43	2.45	--	.99	--	--
2.0	1.74	1.75	--	.99	--	--
3.0	.99	1.00	--	.99	--	--

Set 2

Discrimination Parameters: $\bar{a}$ = .59 , Range = .40; .39 to .79 .
Guessing Parameters : $\bar{c}$ = .00 , Range = .00 .

Ability	$I[\theta,x_1]$	$I[\theta,x_2]$	$I[\theta,x_3]$	$Eff[\theta,x_1]$	$Eff[\theta,x_2]$	$Eff[\theta,x_1,x_2]$
-3.0	.94	.97	--	.97	--	--
-2.0	1.80	1.85	--	.97	--	--
-1.0	2.65	2.74	--	.97	--	--
0.0	2.80	2.91	--	.96	--	--
1.0	2.32	2.39	--	.97	--	--
2.0	1.64	1.69	--	.97	--	--
3.0	.95	.97	--	.98	--	--

Set 3

Discrimination Parameters: $\bar{a}$ = .59 , Range = .80; .19 to .99 .
Guessing Parameters : $\bar{c}$ = .00 , Range = .00 .

Ability	$I[\theta,x_1]$	$I[\theta,x_2]$	$I[\theta,x_3]$	$Eff[\theta,x_1]$	$Eff[\theta,x_2]$	$Eff[\theta,x_1,x_2]$
-3.0	.75	.87	--	.86	--	--
-2.0	1.55	1.79	--	.87	--	--
-1.0	2.59	2.98	--	.87	--	--
0.0	2.73	3.13	--	.87	--	--
1.0	2.01	2.29	--	.87	--	--
2.0	1.34	1.52	--	.88	--	--
3.0	.77	.87	--	.87	--	--

TABLE 1 (Cont'd)

Set 4

Discrimination Parameters: $\bar{a}$ = .59 , Range = .20; .49 to .69 .
Guessing Parameters : $\bar{c}$ = .10 , Range = .20; .00 to .20 .

Ability	$I[\theta,x_1]$	$I[\theta,x_2]$	$I[\theta,x_3]$	$Eff[\theta,x_1]$	$Eff[\theta,x_2]$	$Eff[\theta,x_1,x_2]$
-3.0	.40	.39	.54	.74	.73	1.01
-2.0	1.06	1.07	1.22	.87	.87	.99
-1.0	1.84	1.86	1.95	.94	.95	.99
0.0	2.19	2.23	2.25	.97	.99	.99
1.0	2.03	2.05	2.05	.99	1.00	.99
2.0	1.53	1.54	1.53	.99	1.00	.99
3.0	.89	.90	.90	.99	1.00	.99

Set 5

Discrimination Parameters: $\bar{a}$ = .59 , Range = .40; .39 to .79 .
Guessing Parameters : $\bar{c}$ = .10 , Range = .20; .00 to .20 .

Ability	$I[\theta,x_1]$	$I[\theta,x_2]$	$I[\theta,x_3]$	$Eff[\theta,x_1]$	$Eff[\theta,x_2]$	$Eff[\theta,x_1,x_2]$
-3.0	.39	.37	.53	.73	.70	1.04
-2.0	1.04	1.05	1.21	.86	.87	.99
-1.0	1.86	1.92	2.02	.92	.95	.97
0.0	2.18	2.27	2.30	.95	.99	.96
1.0	1.93	2.00	2.01	.96	1.00	.96
2.0	1.44	1.48	1.48	.97	1.00	.97
3.0	.85	.87	.87	.98	1.00	.98

Set 6

Discrimination Parameters: $\bar{a}$ = .59 , Range = .80; .19 to .99 .
Guessing Parameters : $\bar{c}$ = .10 , Range = .20; .00 to .20 .

Ability	$I[\theta,x_1]$	$I[\theta,x_2]$	$I[\theta,x_3]$	$Eff[\theta,x_1]$	$Eff[\theta,x_2]$	$Eff[\theta,x_1,x_2]$
-3.0	.34	.29	.48	.70	.60	1.16
-2.0	.92	.94	1.14	.80	.82	.97
-1.0	1.83	2.06	2.17	.84	.95	.89
0.0	2.12	2.45	2.48	.85	.99	.86
1.0	1.67	1.92	1.93	.86	1.00	.87
2.0	1.18	1.34	1.34	.88	1.00	.88
3.0	.69	.79	.79	.87	1.00	.87

TABLE 1 (Cont'd)

Set 7

Discrimination Parameters: $\bar{a}$ = .59 , Range = .20; .49 to .69 .
Guessing Parameters : $\bar{c}$ = .20 , Range = .00 .

Ability	$I[\theta,x_1]$	$I[\theta,x_2]$	$I[\theta,x_3]$	$Eff[\theta,x_1]$	$Eff[\theta,x_2]$	$Eff[\theta,x_1,x_2]$
-3.0	.22	.22	.32	.68	.68	1.01
-2.0	.69	.69	.87	.79	.80	.99
-1.0	1.33	1.35	1.53	.87	.89	.98
0.0	1.70	1.72	1.82	.93	.95	.98
1.0	1.64	1.66	1.69	.97	.98	.99
2.0	1.28	1.29	1.29	.99	1.00	.99
3.0	.76	.77	.77	.99	1.00	.99

Set 8

Discrimination Parameters: $\bar{a}$ = .59 , Range = .40; .39 to .79 .
Guessing Parameters : $\bar{c}$ = .20 , Range = .00 .

Ability	$I[\theta,x_1]$	$I[\theta,x_2]$	$I[\theta,x_3]$	$Eff[\theta,x_1]$	$Eff[\theta,x_2]$	$Eff[\theta,x_1,x_2]$
-3.0	.22	.21	.32	.68	.65	1.05
-2.0	.67	.67	.85	.79	.79	1.00
-1.0	1.35	1.40	1.58	.85	.89	.96
0.0	1.69	1.78	1.88	.90	.95	.95
1.0	1.56	1.62	1.65	.95	.98	.96
2.0	1.20	1.23	1.24	.97	1.00	.97
3.0	.73	.74	.74	.98	1.00	.98

Set 9

Discrimination Parameters: $\bar{a}$ = .59 , Range = .80; .19 to .99 .
Guessing Parameters : $\bar{c}$ = .20 , Range = .00 .

Ability	$I[\theta,x_1]$	$I[\theta,x_2]$	$I[\theta,x_3]$	$Eff[\theta,x_1]$	$Eff[\theta,x_2]$	$Eff[\theta,x_1,x_2]$
-3.0	.19	.16	.29	.66	.54	1.23
-2.0	.59	.59	.78	.76	.76	1.00
-1.0	1.33	1.51	1.69	.79	.89	.88
0.0	1.65	1.94	2.06	.80	.95	.85
1.0	1.34	1.56	1.59	.85	.98	.86
2.0	.97	1.11	1.11	.87	1.00	.88
3.0	.58	.67	.67	.87	1.00	.87

TABLE 1 (Cont'd)

Set 10

Discrimination Parameters: $\bar{a}$ = .59 , Range = .00 .
Guessing Parameters : $\bar{c}$ = .10 , Range = .20; .00 to .20 .

Ability	$I[\theta,x_1]$	$I[\theta,x_2]$	$I[\theta,x_3]$	$Eff[\theta,x_1]$	$Eff[\theta,x_2]$	$Eff[\theta,x_1,x_2]$
-3.0	.40	--	.55	.74	--	--
-2.0	1.06	--	1.21	.87	--	--
-1.0	1.80	--	1.90	.95	--	--
0.0	2.19	--	2.22	.98	--	--
1.0	2.09	--	2.10	1.00	--	--
2.0	1.58	--	1.58	1.00	--	--
3.0	.91	--	.91	1.00	--	--

Set 11

Discrimination Parameters: $\bar{a}$ = .59 , Range = .00 .
Guessing Parameters : $\bar{c}$ = .20 , Range = .00 .

Ability	$I[\theta,x_1]$	$I[\theta,x_2]$	$I[\theta,x_3]$	$Eff[\theta,x_1]$	$Eff[\theta,x_2]$	$Eff[\theta,x_1,x_2]$
-3.0	.22	--	.33	.68	--	--
-2.0	.68	--	.87	.79	--	--
-1.0	1.30	--	1.48	.88	--	--
0.0	1.69	--	1.79	.95	--	--
1.0	1.69	--	1.72	.98	--	--
2.0	1.33	--	1.33	1.00	--	--
3.0	.78	--	.78	1.00	--	--

in some testing applications.

The results for efficiency may be summarized as follows: When there is no guessing (i.e. $\bar{c} = 0$), the efficiency of a scoring system using unit weights remains high (over 95%) until the range of the distribution of discrimination parameters becomes large (0.80 in this study). Moreover, efficiency is relatively constant across different levels of ability. When guessing is introduced, this picture changes dramatically. Then, at low ability levels the efficiency of scoring systems $\underline{x}_1$ or $\underline{x}_2$ is markedly reduced, independently of the magnitude of the range of the distribution of discrimination parameters. Of course, as this range increases, the efficiency of $\underline{x}_1$ and $\underline{x}_2$ decreases, again most noticeably at the low ability levels. Indeed, even with a maximum range of the distribution of discrimination parameters (0.80), $\underline{x}_2$ still provides very efficient estimates of ability for examinees with high ability. Under the same circumstances, $\underline{x}_1$ has considerably reduced efficiency.

On the basis of these results, it appears that when a test is being used to estimate ability across a broad range of ability levels and when guessing is a factor in test performance, the scoring system of the three-parameter model is to be preferred. On the other hand, if only high ability examinees are of interest, then even in the presence of guessing, the scoring system of the two-parameter model provides acceptable ability estimates no matter how wide the range of the distribution parameters becomes within the limits studied here. Unit scoring weights, that is the scoring system of the one-parameter (Rasch) model appears to provide efficient estimates of ability when there is little or no guessing and when the range of the distribution of discrimination parameters is fairly small.

References

Birnbaum, A. Efficient design and use of tests of a mental ability for various decision-making problems. Series Report No. 58-16, Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1957.

Birnbaum, A. On the estimation of mental ability. Series Report No. 15, Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958a.

Birnbaum, A. Further considerations of efficiency in tests of a mental ability. Technical Report No. 17. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958b.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical Theories of Mental Test Scores, Reading, Mass: Addison-Wesley, 1968.

Hambleton, R. K. An empirical investigation of the Rasch test theory model. Unpublished doctoral dissertation, University of Toronto, 1969.

Hambleton, R. K. Computation of information curves and efficiency for logistic test models, Fortran IV program for the CDC 3600 Computer. Centre for Educational Research, School of Education, University of Massachusetts, Amherst, Mass., 1970.

Lord, F. M. A theory of test scores. Psychometric Monograph, No. 7, 1952.

Lord, F. M. An analysis of the verbal scholastic aptitude test using Birnbaum's three parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

McDonald, R. P. Nonlinear factor analysis. Psychometric Monograph, No. 15, 1967.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedogogishe Institut, 1960.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Ross, J. An empirical study of a logistic mental test model. Psychometrika, 1966, 31, 325-340.